Data Cleaning and Processing Procedure

Step 1: Load the Dataset

1. Import necessary libraries such as pandas and numpy.
2. Load the dataset using pd.read_csv().

Step 2: Initial Data Examination

1. Display the first few rows of the dataset using df.head().
2. Check the data types of each column using df.dtypes.
3. Identify missing values using df.isnull().sum().

Step 3: Handle Missing Values

1. **Categorical Columns**: Fill missing values with the mode (most frequent value) of the column.
2. **Numerical Columns**: Fill missing values with the median of the column.

Step 4: Remove Outliers

1. Use the Interquartile Range (IQR) method to identify outliers.
    - Calculate Q1 (25th percentile) and Q3 (75th percentile) for each numerical column.
    - Compute IQR = Q3 - Q1.
    - Determine lower bound as Q1 - 1.5 * IQR and upper bound as Q3 + 1.5 * IQR.
2. Filter out rows where numerical column values fall outside the lower and upper bounds.

Step 5: Standardize Data Formats

1. Convert date columns to datetime format using pd.to_datetime().

Step 6: Correct Errors and Remove Duplicates

1. Identify and correct any known errors in the data.
2. Remove duplicate rows using df.drop_duplicates().

Step 7: Feature Engineering

1. Create new features based on existing columns to enhance predictive power.
    - Example: If relevant columns are present, create a new column Yards_per_Second by dividing Yards.Gained by TimeSecs.

Step 8: Data Transformation

1. Normalize or scale numerical columns to bring them onto a similar scale.
    - Use StandardScaler or MinMaxScaler from sklearn.preprocessing to standardize numerical columns.

Step 9: Data Exploration

1. Visualize the distribution of numerical columns using histograms or KDE plots.
2. Explore relationships between variables using pair plots or correlation matrices.